REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	February 10, 2005	Final progress (May 15, 2001 - Dec. 31, 2004)

4. TITLE AND SUBTITLE
Decision trees for prediction and data mining

5. FUNDING NUMBERS
DAAD19-01-1-0586

6. AUTHOR(S)
Wei-Yin Loh

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Wisconsin, 1300 University Avenue, Madison, WI 53706

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
41899.6-MA

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12 a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Three tree-structured algorithms are developed: GUIDE, CRUISE, and LOTUS. GUIDE is an algorithm for constructing least-squares, quantile, and Poisson regression trees. It can fit piecewise constant, piecewise polynomial, and piecewise multiple linear models -- with and without stepwise variable selection. CRUISE is a multiple-branching classification tree algorithm sensitive to pairwise variable interactions at each node. It can fit a single class label or fit the best two-variable linear discriminant model to each leaf node. LOTUS is a logistic regression tree algorithm. It can fit either simple linear or multiple linear logistic regression models at the leaf nodes. All three algorithms are unique in being practically free of variable selection bias at the splits. Empirical comparisons with a variety of real datasets show that the algorithms possess both high prediction accuracy and fast computational speed. Computer software for the algorithms are available for free download from the internet for the Windows, Linux, and (in the case of GUIDE) Macintosh platforms.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Regression trees, unbiased variable selection, classification, logistic regression	7 (including this page)
	16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION ON THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	**UNCLASSIFIED**	**UNCLASSIFIED**	**UL**

NSN 7540-01-280-5500

Standard Form 298 (Rev.2-89)
Prescribed by ANSI Std. 239-18
298-102

Enclosure 1

1 Problems studied

The main focus of the research is on the accuracy, interpretability, and visualizability of classification and regression trees. This is partly motivated by the recent interest, within the statistics and data mining communities, in averaging across ensembles of trees to increase prediction accuracy (Breiman 1996, Breiman 1998). A serious disadvantage of ensemble techniques is the impracticality of simultaneously interpreting more than a very small number of trees. This ironic, as research in tree-structured methods was originally motivated by the desire for an interpretable alternative to standard methods such as multiple linear regression and neural networks.

Another problem with most tree construction algorithms is that their variable selection methods are biased towards choosing some types of variables over others. As a result, conclusions drawn from such trees can be, and often are, wrong.

A primary goal of our project is to design algorithms for trees of sufficient complexity (not in terms of size but in the type of splits and node models) that the accuracy of a single tree is comparable to that of an ensemble of trees. In addition, the trees are free of variable selection bias. Another important practical goal is to implement the algorithms into high-quality computer software for Windows, Linux, Macintosh and other operating systems.

2 Summary of important results

2.1 Linear regression trees

The most progress is made in this area, because the PI is solely and totally responsible for the design and implementation of the GUIDE algorithm. At the time that this report is written, GUIDE has already far out-paced all other regression tree software, including well-known ones such as CART (Breiman, Friedman, Olshen and Stone 1984) and M5 (Quinlan 1992). With the exception of the lesser-known RT (Torgo 1999), the other methods are exclusively designed for piecewise-constant least squares regression. GUIDE can produce piecewise-polynomial and piecewise-multiple linear (including stepwise regression) models as well. Further, GUIDE is the only algorithm that can fit quantile and Poisson regression models. But the most important and unique feature of GUIDE is that its variable selection procedure for splitting each

node is bias-corrected. This is a very tricky problem that no other regression tree algorithm has been able to solve. Without bias-correction, a tree model can be worse than useless for interpretation because of the potential for incorrect inferences.

The GUIDE bias correction approach for least squares and Poisson regression is explained in article [4]; it is extended to quantile regression in [6] (article numbers refer to those in Section 3.1). Two other manuscripts have been submitted for publication. One demonstrates how GUIDE can be used to visualize high-dimensional datasets. It also contains the results of a large-scale empirical study showing GUIDE to have as good prediction accuracy as the best tree or non-tree regression algorithms, the latter including spline models such as GAM (Hastie and Tibshirani 1990) and MARS (Friedman 1991), and rule-based models from the computer science literature. Another manuscript ilustrates the advantages of GUIDE in fitting models to classical factorial experiments where the sample sizes are small. The titles of the articles are listed in Section 3.2.

2.2 Logistic regression trees

The main appeal of a logistic regression tree over a classification tree is its ability to classify as well as to attach a probability to each subject. The latter allows the ranking of subjects which is important to the service industry, for example. Thus, using a logistic regression tree, a company can determine the top 20% of its customers most likely to be dissatisfied with its service/product, and try to improve its product design and services to achieve higher customer satisfaction.

Three of the PI's students have written PhD theses on this problem (Lo 1993, Potter 1998, Chan 2000). The latest algorithm is called LOTUS. Like GUIDE, its distinguishing feature is that it has negligible variable selection bias. Further, it is relatively computation inexpensive and can handle both numerical and categorical covariates. Finally, it is flexible in the type of linear model used and has a built-in mechanism to handle missing values. Evaluations based on real and simulated data show that LOTUS performs well in most situations. The results are reported in articles [9] and [11] in Section 3.1.

2.3 Classification trees

The QUEST algorithm (Loh and Shih 1997) was stable during the period of this research. The software received a small number of bug fixes. QUEST is the PI's most popular software, as measured by the hundred or so hits its website receives each week. This is probably due to its maturity (the first version was released more than ten years ago) and to the adoption of its basic algorithm by commercial publishers SPSS Inc., and StatSoft.

The most significant new development in this area is the CRUISE algorithm, which extends the unbiasedness of QUEST in several directions. First, it splits each node of a tree into as many branches as the number of values taken by the response variable. When the dataset is very large, this has the advantage of producing a shorter, and hence more comprehensible, tree. CHAID (Kass 1980) is the only other algorithm with non-binary splits. But CHAID is otherwise quite different, because the number of splits is fixed at ten for each ordered predictor variable and the tree is not pruned.

The other unique feature is CRUISE's sensitivity to local interactions between pairs of variables. All other classification tree algorithms concentrate only on one variable at a time. Two papers on CRUISE were published during the period of the grant. They are [2] and [7] in Section 3.1. A third paper [8] gives an application of CRUISE to a problem in construction engineering.

3 Publications and technical reports

3.1 Peer-reviewed journals

1. Asymptotic theory for Box-Cox transformations in linear models (with K. Cho, I. Yeo, and R. A. Johnson). *Statistics and Probability Letters*, 2001, **51**, 337–343.

2. Prediction interval estimation in transformed linear models (with K. Cho, I. Yeo, and R. A. Johnson). *Statistics and Probability Letters*, 2001, **51**, 345–350.

3. Classification trees with unbiased multiway splits (with H. Kim). *Journal of the American Statistical Association*, 2001, **96**, 589–604.

4. Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 2002, **12**, 361–386.

5. A framework for measuring differences in data characteristics (with V. Ganti, J. Gehrke and R. Ramakrishnan). *Journal of Computer and System Sciences* 2002, **64**, 542–578.

6. Nonparametric estimation of conditional quantiles using quantile regression trees (with P. Chaudhuri). *Bernoulli*, 2002, **8**, 561–576.

7. Classification trees with bivariate linear discriminant node models (with H. Kim). *Journal of Computational and Graphical Statistics*, 2003, **12**, 512–530.

8. Decision tree approach to classify and quantify cumulative impact of change orders on productivity (with M. J. Lee and A. S. Hanna). *Journal of Computing in Civil Engineering*, 2004, **18**, 132–144.

9. LOTUS: An algorithm for building accurate and comprehensible logistic regression trees (with K-Y Chan). *Journal of Computational and Graphical Statistics*, 2004, **13**, 826–852.

10. Box-Cox transformations. Book chapter in *Encyclopedia of Statistical Sciences*, 2nd edition, Wiley. In press.

11. Logistic regression tree analysis. Book chapter in *Handbook of Engineering Statistics*, Springer. In press.

3.2 Manuscripts submitted

1. A visualizable and interpretable regression model with good prediction power (with H. Kim, Y.-S. Shih, and P. Chaudhuri). Submitted to *IIE Transactions Special Issue on Data Mining*.

2. Regression tree models for designed experiments. Submitted to *Proceedings of the Second Lehmann Symposium*.

4 Scientific personnel and advanced degrees earned

Two PhD students were supported as research assistants at various times during the grant period:

1. Hyungjun Cho

2. Qinghua Song

Hyungjun Cho received his PhD degree under the PI's supervision in 2002 (Cho 2002). He is currently a postdoctoral fellow at the University of Virginia. Qinghua Song is expected to complete his degree in the next twelve months.

5 Software developed

The executable binaries (for Linux, Windows, and others) of the following algorithms are being distributed free from the PI's website `http://www.stat.wisc.edu/~loh/`.

1. CRUISE classification tree

2. GUIDE regression tree

3. LOTUS logistic regression tree

4. QUEST classification tree

Bibliography

Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**: 123–140.

Breiman, L. (1998). Arcing classifiers (with discussion), *Annals of Statistics* **26**: 801–849.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, Wadsworth, Belmont.

Chan, K. Y. (2000). *Logistic Regression Trees*, PhD thesis, Department of Statistics, University of Wisconsin.

Cho, H. (2002). *Tree-structured regression modeling for censored data*, PhD thesis, Department of Statistics, University of Wisconsin.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion), *Annals of Statistics* **19**: 1–141.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman and Hall, London.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data, *Applied Statistics* **29**: 119–127.

Lo, W.-D. (1993). *Logistic Regression Trees*, PhD thesis, Department of Statistics, University of Wisconsin, Madison.

Loh, W.-Y. and Shih, Y.-S. (1997). Split selection methods for classification trees, *Statistica Sinica* **7**: 815–840.

Potter, D. (1998). *Logistic Regression Trees*, PhD thesis, Department of Statistics, University of Wisconsin.

Quinlan, J. R. (1992). Learning with continuous classes, *5th Australian Joint Conference on Artificial Intelligence*, pp. 343–348.

Torgo, L. (1999). *Inductive Learning of Tree-based Regression Models*, PhD thesis, Department of Computer Science, Faculty of Sciences, University of Porto.